УДК: 004.3

1.11. Совместное использование ресурсов в индустрии 4.0

Балычев С. Ю.[1], Батьковский А. М.[2], Батьковский М. А.[3], Белоусов Ф. А.[2], Неволин И.В.[2]
[1] Финансовый университет при Правительстве РФ, Москва, Россия;
[2] Центральный экономико-математический институт Российской академии наук,
[3] АО «НИЦ «Интелэлектрон», Москва, Россия

Цель настоящего исследования – анализ распределения вычислительных ресурсов, используемых фирмами при осуществлении своей деятельности на основе концепции Индустрии 4.0. Другой уровень организации производства и управления, повышение гибкости и адаптивности бизнес-систем за счет развития автоматизации и обмена данными в постоянном взаимодействии с внешней средой – эти аспекты цифровой экономики влияют на сложность проблемы в контексте. Развитие информационных технологий должно сопровождаться разработкой новых решений для поддержки изменений, с которыми сталкивается отрасль. Поэтому оптимизация распределения вычислительных ресурсов в условиях распространения технологий Индустрии 4.0 приобретает большое научное и практическое значение. Методологический подход, используемый в исследовании, основывается на фундаментальных результатах исследований в области трансформации информационно-коммуникационных технологий, общих принципах теории управления, методах ЭКВ-данных.

1. Введение

1.1. Параллельные вычисления

В связи с растущим спросом на высокопроизводительные вычисления необходимы инструменты и методы для управления очередью задач с учетом возрастающей важности и безопасности. Удовлетворение спроса на вычислительную технику способствует научному прогрессу и промышленному развитию, обеспечивая достижение целей бизнеса, правительства и общества [Batkovskiy, Sudakov, Fomina, 2020]. Ограниченные и дорогостоящие аппаратные ресурсы ставят вопрос о приоритете при планировании вычислений и выделении им вычислительной мощности. Следует принимать во внимание разделение процессов в общей инфраструктуре, а также частные знания, лежащие в основе обрабатываемой задачи. Эти знания относятся к алгоритмам, технике, используемой в вычислениях, и бизнес-контексту конкретной задачи. Все это обеспечивает конкурентное преимущество в индустрии 4.0. Таким образом, распределение ресурсов между задачами принимает форму проблемы с неполной информацией. Мы формулируем это как задачу оптимизации для высокотехнологичной фирмы по обеспечению своевременного выполнения ее разнообразных проектов. Например, это подходит для оборонных предприятий, разрабатывающих программы и планы диверсификации производства.

Индустрия 4.0 предполагает использование высокопроизводительных вычислений и больших данных в процессах проектирования и производства [Tay et al, 2018]. Цифровые близнецы относятся к числу основных объектов и моделей в отрасли. Глубокое понимание этого термина выходит за рамки статистических моделей. Действительно, он основан на системе конечно-разностных уравнений [Боровков, 2018]. Цифровые двойники – компьютерные модели реальных объектов – позволяют инженерам быстро находить успешные дизайнерские решения и использовать максимум доступных возможностей благодаря современным материалам, а также тестировать их на виртуальных стендах. Именно компьютерный компонент позволяет сократить количество прототипов и, как следствие, сократить стадию разработки и быстрее вывести продукт на рынок [Боровков, 2018].

Суперкомпьютерные технологии и обработка больших объемов данных стали ключом к конкурентным преимуществам, а их эффективное использование связано с вопросами архитектуры систем, безопасности, юридическими проблемами и значимыми результатами. Последний пункт понятен: расчеты должны быть максимально приближены к реальности и основываться на модели с однозначной интерпретацией. Например, в случае цифровых двойников для автомобильной промышленности качество моделирования обеспечивает результаты, которые отличаются от физических тестов на несколько процентов [Боровков, 2018].

Юридические проблемы связаны с используемым программным обеспечением и наборами данных. Нелицензионные продукты и несанкционированный доступ к данным создают риски, как минимум, прерывания проекта, а в худшем случае – судебного преследования. Особый вопрос в юридической литературе касается правообладателя результатов моделирования и критериев, удовлетворяющих различным правовым требованиям для защиты [Gervais, 2019].

Концепция "больших данных" – неотъемлемая часть вычислительной техники в рамках индустрии 4.0 – характеризует скорость, разнообразие и объем данных. Можно кратко выразить основную идею больших данных с некоторым упрощением, указав на распределенную среду хранения и обработки: когда аппаратные и программные возможности выходят за рамки одного компьютера, задействуются технологии распараллеливания между отдельными устройствами [International Organization for Stand-

ardization (ISO) and International Electrotechnical Commission (IEC), 2019]. Более того, параллелизм возникает не только в контексте хранения данных, но и в вычислениях: физические задачи, решаемые на пространственной сетке, очень естественны для параллельных вычислений [Saad, 2003]. Таким образом, когда дело доходит до реализации, вопрос архитектуры относится к диапазону устройств, их количеству, аппаратным настройкам, а также используемому программному обеспечению [ISO & IEC, 2020]. Инструмент для HPC может быть собран либо на самостоятельно установленной инфраструктуре фирмы, либо в облаке, арендованном у внешнего поставщика. Термины "самостоятельно установленный" или "внешний поставщик" относятся к организации – субъекту высокотехнологичной отрасли. Действительно, самостоятельно установленная инфраструктура является активом организации, в то время как облачный провайдер является внешним по отношению к организации с точки зрения прав собственности. Профессиональное сообщество уделяет особое внимание архитектуре систем, использующих большие данные, по крайней мере, по двум причинам. Первая – это разработка стандартов для таких систем, которые охватывали бы основные концепции проектирования, верификации, безопасности и обработки данных. Вторая – это выбор коммерческого продукта, сделанный фирмой в соответствии с задачами обработки больших данных.

1.2. Ответственность за инфраструктуру

Собственные серверы и системы хранения данных позволяют компании устанавливать четкие и прозрачные правила безопасности для пользователей и супервизоров: выбирать подходящее оборудование, управлять установленным программным обеспечением и сетевыми подключениями, определять политику доступа для узлов кластера и т.д. Но недостатком является то, что вся ответственность за происходящее лежит на фирме: она самостоятельно несет все расходы по функционированию кластера, нанимает персонал соответствующей квалификации, приобретает оборудование для обеспечения безопасности и т.д. Облачные технологии, в свою очередь, обещают гибкость масштабирования (емкость может быть быстро увеличена или уменьшена), оплату только за использованные ресурсы. Однако они полагаются на доверие к поставщику услуг с точки зрения безопасности и отказоустойчивости системы. Но, учитывая инфраструктуру за пределами фирмы, пользователь не может быть уверен в том, что его действия не контролируются третьей стороной. Более того, внешний контроль весьма вероятен, поскольку ведение журнала является неотъемлемой частью безопасности [Luo et al, 2019]. В дополнение к внешнему контролю человек раскрывает третьей стороне (поставщику услуг – владельцу вычислительного кластера) план использования ресурсов – требования внутренних команд, разрабатывающих определенные проекты.

В случае регулярной работы с компьютерными моделями самоустанавливающаяся инфраструктура становится разумной. Способность внедрить соответствующую политику безопасности – это всего лишь один из факторов. Однако иногда это становится необходимым, как, например, в оборонной промышленности. Есть и другие аргументы в пользу этого выбора. Анализ показывает, что облачные вычисления обеспечивают более низкую общую стоимость владения в течение короткого периода времени и при низкой потребности в вычислениях [De Alfonso et al, 2013; Filiopoulou, 2020; Nikolaou et al, 2019]. Если расчеты и обработка больших объемов данных входят в повседневную деятельность фирмы, дешевле установить собственную инфраструктуру. Ответственность за внешние атаки и отказоустойчивость системы ложится на фирму. Например, в случае физических вычислений, которые не имеют коммерческой ценности, наиболее типичные атаки направлены на отключение кластера – DDOS-атаки или перегрузку сети, – что, конечно, затрудняет работу и может привести к очистке аппаратной памяти, что влечет за собой потерю результатов текущих вычислений, но не приводит к утечке данных.

Независимо от выбора инфраструктуры, необходимо обеспечить изоляцию процессов, инициируемых отдельными пользователями, друг от друга, а также их хранение. Это хорошо известная техническая проблема, которая особенно актуальна в случае распределенных ресурсов. Фактически, возникает проблема безопасности в широком смысле: как защитить данные от злоумышленников, а также разграничить процессы, которые используют общую инфраструктуру.

1.3. Вклад в работу

Мы рассматриваем фирму, у которой есть портфель задач для высокопроизводительных вычислений. Каждая задача характеризуется потребностью в вычислительных ресурсах, а специфика зависит от различных команд внутри фирмы, каждая из которых разрабатывает отдельный продукт. Запросы команд на ресурсы могут быть искажены – скорее всего, завышены, а реже занижены – объемы вычислительных ядер и памяти. Каковы бы ни были причины этих искажений, их можно объяснить желанием команды обеспечить свое развитие: как в соответствии с внутренними правилами компании, так и по личным соображениям команды. Следующий раздел разъясняет этот момент и показывает, что такое искажение является разумным, поскольку окружающая среда получает неверный сигнал о характере решаемой проблемы и, таким образом, сталкивается с растущими затратами на сбор информации о фирме и ее проектах.

Мы предлагаем алгоритм для поддержки распределения вычислительных ресурсов между различными задачами в условиях неполной информации. Преимущество этого алгоритма заключается в том, что заявителям – командам, ищущим ресурсы, – не нужно раскрывать полную информацию о вли-

нии ядер и памяти на скорость и качество предлагаемого ими решения. Соглашение заключается после реакции заявителей на предложение центрального планировщика. Теория гарантирует сходимость алгоритма и равновесие по Парето при естественных ограничениях.

2. Материалы и методы

Представьте себе фирму, которая разрабатывает множество инженерных проектов. Для тестирования конкретных схем и проектных решений фирма использует цифровую технику. С учетом этого путь к конечному продукту в каждом проекте лежит через серию вычислительных экспериментов с различными моделями, различными условиями и ограничениями. Таким образом, каждому проекту необходим определенный набор ИТ-средств, включая вычислительные узлы, центр обработки данных и сетевую инфраструктуру. Мы оставляем используемое программное обеспечение за рамками обсуждения, предполагая, что физическая инфраструктура, а также программы, поддерживающие вычисления, позволяют каждому проекту выполнять все необходимые вычисления, в том числе параллельные. Как правило, у каждой команды есть уникальный поставщик таких ресурсов. На уровне фирмы вопрос заключается в том, как обеспечить проектные группы необходимыми ресурсами в условиях физических и организационных ограничений фирмы, максимизируя выгоды и согласовывая интересы команд и руководства. В своей работе мы стремимся разработать алгоритм, обеспечивающий равновесие в распределении компьютерных ресурсов между проектными командами внутри фирмы. Равновесие, на котором мы фокусируемся, близко к концепции теории игр: в таком состоянии отклонение любого параметра ухудшает доступность ресурсов для определенного проекта.

Решение задачи оптимизации дает точный ответ на этот вопрос. Часто в литературе по оптимальному планированию инфраструктуры проблема рассматривается с точки зрения энергопотребления [Niewiadomska-Szynkiewicz, Arabas, 2018; Singh, Dziurzanski, Indrusiak, 2015]. Хранение данных и вычислительные мощности потребляют огромное количество энергии – это основная статья расходов, в которую также входят коммуникации, техническое обслуживание, администрирование и т.д. Более того, значительная часть этой литературы посвящена динамической балансировке нагрузки, то есть распределению задач на работающем вычислительном кластере, когда задания поступают в режиме реального времени. Оптимизация такого рода основана на специальных датчиках и контроллерах для изменения напряжения на отдельных устройствах. Такие технические меры выходят за рамки данной статьи, и мы упоминаем динамическую оптимизацию, поскольку она также требует оценки вычислительной сложности поступающих задач – оценки спроса на ресурсы. Это важный момент в нашем контексте, поскольку, во-первых, такая оценка может косвенно указывать на используемые алгоритмы и масштаб проблемы. Во-вторых, нет никаких сомнений в возможности такой оценки. Очевидно, что оборудование определенной спецификации тратит характерное время на типичные операции, однако типология довольно ограничена в контексте организации. Перечень программного обеспечения для моделирования физических процессов часто ограничен: дешевле собрать необходимую конфигурацию из коммерческих пакетов со встроенными алгоритмами, чем разрабатывать новый продукт для конкретной задачи. Но этот программный компонент также известен внутри организации. Другими словами, благодаря популярности используемых инструментов – вычислительного оборудования и программного обеспечения – можно строить догадки о характере решаемой проблемы. Точные параметры часто являются частной информацией, которая остается секретной независимо от того, введен ли в отношении них соответствующий правовой режим или нет. Например, патенты на способы приготовления смесей, на компоненты из сплавов оперируют диапазоном значений в их описании и формуле изобретения. Причина кроется не только в более широком охвате возможных реализаций продукта, но и в сокрытии точных значений, полученных в ходе длительных исследований и многих дорогостоящих экспериментов. Поскольку патент является документом, относящимся к конкретной стране, возможно легальное использование изобретения в юрисдикции, не имеющей правовой охраны. Отсутствие точных характеристик в тексте патента оставляет потенциальному пользователю мало вариантов: нужно либо самостоятельно провести серию экспериментов и скорректировать значения, либо обратиться за помощью к изобретателю. Во втором случае лицензионное соглашение заключается и действует даже на территории, где изобретение не охраняется законом!

Сокрытие подробностей о решаемых задачах даже от команд и группировок внутри фирмы, не говоря уже о неразглашении информации за пределами фирмы, соответствует протоколам информационной безопасности. В настоящее время социальная инженерия занимает значительное место в планировании внешних атак. Это включает в себя поиск инсайдера внутри организации и сбор информации об атакованной инфраструктуре. Среди прочего, например, количество и местоположение серверов – это информация, которая должна быть скрыта для обеспечения отказоустойчивости системы. С точки зрения коммерческих интересов и конкурентных преимуществ, квалификация сотрудников, используемые алгоритмы и разработанные компьютерные модели имеют высокую ценность. Организационные меры по предотвращению утечки передовых разработок включают режим коммерческой тайны с ограничением доступа лиц к соответствующей информации. В то же время даже сигнал о характере работы – запрашиваемых ресурсах и рабочем времени – может стать поводом для пристального внимания к проекту и его участникам.

Это, с одной стороны, необходимые ресурсы, полученные в результате решения задачи оптимизации, а с другой стороны, они рассматриваются как сигнал для злоумышленников и, следовательно, не должны разглашаться. Учитывая это, для достижения вышеупомянутой цели работы – разработки алгоритма, приводящего к равновесному распределению ресурсов, – необходимо решить следующие задачи:

1) описать принятие проектной командой решения о раскрытии заявки на вычислительные ресурсы;
2) описать процедуру согласования стимулов в условиях неполной информации, включая преднамеренные искажения;
3) обеспечить равновесие результирующего распределения;
4) представить решение вышеперечисленных задач в виде формального алгоритма.

Методологическую основу работы составляют методы оптимизации. Согласование стимулов в условиях неполной информации – хорошо известная экономическая проблема, которая могла бы быть решена без участия человека при наличии современных технологий – с помощью компьютерного посредника [Nevolin, Kozyrev, 2014; Zhang, Zhu, 2019]. Методы решения включают двойные аукционы [Ba, Stallaert, Whinston, 2001], последовательные уступки [Гольштейн, Борисова, Дубсон, 1990] и методы из теории переговоров [Ehtamo, Kettunen, Hämäläinen, 2001], предлагающие участникам торгов обменять один ресурс на другой с улучшением полезности.

Акцент на компьютерном посреднике, с одной стороны, обеспечивает объективность принимаемого решения с учетом лежащей в его основе формальной процедуры. С другой стороны, машина обеспечивает довольно быстрый пересчет переговорного набора из ограниченного, хотя и довольно большого числа комбинаций. Двойные аукционы побуждают команды сообщать правдивые оценки и могут потребовать внешних ресурсов для обеспечения совместимости стимулов [Ba, Stallaert, Whinston, 2001] или штрафов, аналогичных основному налогу Кларка [Detering, 2001]. Это не всегда уместно и несовместимо с рассматриваемым контекстом – распределением уже имеющихся вычислительных ресурсов без дополнительных обновлений. Метод последовательных уступок требует пропорционального обмена ресурсами. Таким образом, дополнительно требуется процедура направленного поиска. Согласно теории переговоров, можно оптимизировать усилия для достижения консенсусного подхода, и поэтому эта теория выглядит наиболее многообещающей для формализации процедуры поэтапного достижения результата.

Мы относим наш метод к последнему подходу. Предполагается, что каждая команда внутри фирмы решает проблему оптимизации: она стремится выполнять вычисления требуемого качества с учетом доступного количества ядер и компьютерной памяти. Конечно, общий объем ресурсов фирмы, доступных для одной команды, позволяет ей создать более детализированного цифрового двойника и, следовательно, получить результат более высокого качества. Но монополизация ресурсов привела бы к неудачам в других проектах. Таким образом, ограничения в задаче оптимизации команды должны учитывать спрос со стороны их коллег. Однако, как упоминалось выше, каждая команда излагает свое требование с искаженными показателями. При формализации алгоритма мы предполагаем, что проблема каждого кандидата влечет за собой ее собственные ограничения при построении цифрового двойника, и взаимосвязь между всеми проблемами задается уравнением баланса.

3. Результаты

Введем обозначения для формализации алгоритма распределения компьютерных ресурсов между вычислительными задачами, при этом каждая задача будет выполняться одной командой внутри фирмы. Пусть фирма разрабатывает J продуктов (или исследовательских проектов) с J вычислительными задачами – по одной на продукт (или проект). Задания пронумерованы индексом j. Каждая команда стремится завершить свои расчеты как можно скорее. А спрос на компьютерные ресурсы состоит из множества {t, CP, MP}, где t – ожидаемое время вычисления, CP означает вычислительные ядра, а MP – память компьютера. При планировании ресурсов команда стремится обеспечить качество расчетов (точность аппроксимации, детализацию при разбиении на разделы и т.д.) выше определенного порогового значения q. Ресурсы также имеют естественную нижнюю границу: доступная память компьютера не должна быть ниже порогового значения m. В противном случае компьютер или кластер не смогли бы работать с переменными минимальной размерности. Также естественно предположить, что качество вычисления φ является монотонной выпуклой функцией, которая увеличивается с увеличением доступных ресурсов. Эта функция, как и функция времени на решение задачи, является частным знанием команды, что позволяет классифицировать проблему распределения компьютерных ресурсов как оптимизацию в условиях неполной информации. Итак, у нас есть J задач выпуклого программирования:

$$t_j(CP_j, MP_j) \to min \quad (1)$$

s.t.

$$\varphi_j(CP_j, MP_j) \geq q_j \quad (2)$$

$$MP_j \geq m_j \quad (3)$$

$$CP_j \geq 0 \quad (4)$$

Эти задачи связаны с двумя ограничениями:

$\sum_{j=1}^{J} CP_j = \overline{CP}$ (5)

$\sum_{j=1}^{J} MP_j = \overline{MP}$ (6)

Мы следуем работе А.Н. Козырева [Козырев, 1975] по разработке алгоритма выравнивания стимулов в условиях неполной информации. По своей сути это итеративный процесс. На каждом этапе командам предлагается решение, и их реакция рассматривается как встречное предложение ресурсов. Вначале нужна отправная точка – пусть это будет набор $\{{CP_1}^0, \dots, {CP_J}^0, {MP_1}^0, \dots, {MP_J}^0\}$. В ответ каждая команда j выдвигает свой ответ $\{{CP_j}^1, {MP_j}^1\}$ в командный центр, или посреднику. Этот ответ на самом деле является аппроксимацией субдифференциала функции полезности $t_j(CP_j, MP_j)$. Получив ответы, посредник (программное обеспечение могло бы выполнять эту роль благодаря формализации алгоритма) перераспределяет доступные ресурсы $\{\overline{CP}, \overline{MP}\}$ между командами в соответствии с их предложениями. В результате до команд доводится набор $\{{CP_1}^2, \dots, {CP_J}^2, {MP_1}^2, \dots, {MP_J}^2\}$, и все отвечают новым набором $\{{CP_j}^3, {MP_j}^3\}$. Это продолжается до тех пор, пока ответы команд не будут отличаться от заявленного набора на небольшое предопределенное значение. А.Н. Козырев доказал, что для линейной функции $tt_j(CP_j, MP_j)$ и $\varphi_j(CP_j, MP_j)$ алгоритм сходится и обеспечивает Парето-оптимальность решения. Фактически, алгоритм также сходится в случае монотонных неубывающих функций.

При заданных субдифференциалах алгоритм работает следующим образом:

1) Посредник сообщает о текущем распределении ресурсов.
2) Посредник получает ответы от команд, каждый из которых является приближением субградиента в текущей точке.
3) С учетом субградиентов вычисляется луч с вектором направления, равным сумме субградиентов команд.
4) Для каждого ответа посредник вычисляет вектор обмена как проекцию на ортогональное построенному выше лучу подпространство.
5) С учетом неполноты информации коэффициенты масштабирования рассчитываются таким образом, чтобы новое распределение не выходило за пределы диапазона допустимых значений. Минимальный коэффициент выбирается при достижении одной из команд его предела с точки зрения качества или компьютерной памяти.
6) С учетом результатов обмена ресурсами между командами рассчитывается новое распределение.
7) Повторяются шаги (1) – (6).

В нашей работе мы предполагаем, что проектная команда является рациональным агентом, который оптимизирует свою полезность, решая задачу, аналогичную уравнениям (1)-(4). Это предположение в соответствии с задачей 1 данного исследования необходимо для доказательства равновесия – задачи 3 исследования. Однако даже при использовании других механизмов, лежащих в основе решения, сходимость процедуры гарантируется в соответствии с упорядоченными предпочтениями агентов.

Следует отметить, что компьютерный посредник важен для согласования стимулов (задача 2 исследования), поскольку именно машина вычисляет обмен ресурсами и связывает участников. Его вторая роль заключается в том, чтобы скрывать заявку каждой команды от остальных участников переговоров. Эта звездообразная топология, когда все команды подключены к центру управления, имеет преимущество перед технологиями распределенного реестра в этом аспекте. Ни один участник не имеет доступа к другим приложениям, даже в зашифрованном виде.

4. Обсуждение

Литература по оптимизации вычислительных ресурсов часто рассматривает эту тему с точки зрения энергопотребления [Niewiadomska-Szynkiewicz, Arabas, 2018; Singh, Dziurzanski, Indrusiak, 2015]. Такие работы наиболее актуальны для центров обработки данных и поставщиков облачных услуг, поскольку в этих случаях решающее значение имеют затраты на обслуживание инфраструктуры, обеспечивающей поток работ. В то же время избыточный спрос на них не является проблемой: принятые процедуры просто “отключают” на некоторое время простаивающие мощности. В модельной ситуации, на которой мы фокусируемся, важно как можно точнее выявить предпочтения. Ресурсы фирмы значительно более ограничены, чем у поставщиков услуг. Поэтому важно максимально эффективно использовать существующую инфраструктуру, избегая простоев проектных команд.

Очевидно, что, следуя алгоритму оптимизации, какая-то команда (или даже несколько команд) могла бы располагать избыточными ресурсами – больше, чем на самом деле необходимо для вычислений. Это связано с уравнениями баланса (5)-(6): все вычислительные ядра и вся память должны быть распределены между задачами. Это предполагает две возможности для фирмы. Во-первых, команда получает результат более высокого качества в соответствии с имеющимися ресурсами. Во-вторых, избыточная мощность может быть “отключена” в режиме динамической балансировки нагрузки. То есть наш алгоритм применим к той части архитектуры высокопроизводительных систем, которая

называется Global Computing Resource Manager, в то время как менеджер локальных вычислительных ресурсов отвечает за динамическую балансировку [Niewiadomska-Szynkiewicz, Arabas, 2018].

Наконец, мы должны отметить предположения о монотонности функции времени t и функции качества φ. Алгоритм гарантирует, что каждое новое отчетное распределение будет не хуже предыдущего. Но это происходит только тогда, когда функции ресурса монотонны. Если ответ посреднику противоречит тенденции в обмене ресурсами, следует задуматься об отсутствии представления о форме функций t и φ внутри команды или оппортунистическом поведении команды, стремящейся достичь дополнительных целей, выходящих за рамки заявленных. В любом случае, это повод повнимательнее присмотреться к деятельности такой команды.

5. Заключение

Предлагаемый алгоритм распределения вычислительных ресурсов внутри фирмы относится к индустрии 4.0. Данные и высокопроизводительные вычисления играют все возрастающую роль в экономике. В результате у сотрудников возрастает спрос на вычислительную инфраструктуру, и эта инфраструктура – ограниченный и дорогостоящий ресурс – должна быть распределена между ними. Однако политика безопасности часто ограничивает раскрытие информации, и поэтому руководство фирмы сталкивается с проблемой оптимизации в условиях неполной информации. Решение проблемы предполагает последовательный обмен ресурсами внутри фирмы, при этом каждый участник объявляет только об увеличении или уменьшении объема получаемых ресурсов. Такой подход не исключает динамического балансирования нагрузки. Действительно, это элемент планирования – этап, предшествующий вычислениям на вычислительном кластере.

Применение алгоритма не подходит для случая облачных вычислений, поскольку администратор - поставщик услуг – заинтересован в высоком спросе, а дополнительные условия оттолкнули бы новых пользователей. В то же время новые рабочие места появляются динамично, и стимулы пользователей невозможно согласовать заранее. Остается только встроить новые задания в существующую очередь.

Благодарности

Статья подготовлена при финансовой поддержке Российского научного фонда, грант № 21–78–20001 «Разработка теории и модельных инструментов оптимизации управления диверсификацией оборонного производства в условиях экономического кризиса и растущих угроз национальной безопасности России».

Литература

1. Боровков А.И. (2018) Умные технологии на службе продуктовых программ // Проектный вестник, 2, 32–36.
2. Гольштейн Е.Г., Борисова Э.П., Дубсон М.С. (1990) Диалоговая система анализа многокритериальных задач //Экономика и математические методы, 26 (4).
3. Козырев А.Н. (1975). Оптимизация распределения ресурсов в системе линейных производственных моделей. Оптимизация, 16(33), 62.
4. Ba, S., Stallaert, J., & Whinston, A. B. (2001). Optimal investment in knowledge within a firm using a market mechanism. *Management Science, 47*(9), 1203–1219.
5. Batkovskiy, A.M., Sudakov, V.A. & Fomina, A.V. (2020) Minimization of Information Losses in the Management of Innovation Development in the Digital Economy. *Lecture Notes in Networks and Systems*, 115, 336–343
6. De Alfonso, C., Caballer, M., Alvarruiz, F., & Moltó, G. (2013). An economic and energy-aware analysis of the viability of outsourcing cluster computing to a cloud. *Future Generation Computer Systems, 29*(3), 704-712
7. Detering, D. (2001). *Ökonomie der Medieninhalte: allokative Effizienz und soziale Chancengleichheit in den neuen Medien*. Münster: LIT Verlag.
8. Ehtamo, H., Kettunen, E., & Hämäläinen, R. P. (2001). Searching for joint gains in multi-party negotiations. *European Journal of Operational Research, 130*(1), 54–69.
9. Filiopoulou, E. (2020). *Analysis of Pricing Strategies of Infrastructure as a Service (IaaS)* (PhD dissertation). Athens, Greece: Harokopio University.
10. Gervais, D. (2019). Exploring the interfaces between big data and intellectual property law. *J. Intell. Prop. Info. Tech. & Elec. Com. L.,* 10, 3.
11. Information technology — Big data — Overview and vocabulary. (2019). ISO/IEC 20546:2019 from Feb 2019. Geneva, Switzerland: International Organization for Standardization and International Electrotechnical Commission
12. Information technology — Big data reference architecture — Part 3: Reference architecture. (2020). ISO/IEC 20547-3:2020 from Mar 2020. Geneva, Switzerland: International Organization for Standardization and International Electrotechnical Commission
13. Luo et al, 2019 Luo, Z., Qu, Z., Nguyen, T., Zeng, H., & Lu, Z. (2019). Security of HPC systems: From a log-analyzing perspective. *EAI Endorsed Transactions on Security and Safety, 6*(21), e5.
14. Nevolin, I., & Kozyrev, A. (2014). Developing CRIS module for technology transfer. *Procedia Computer Science*, 33, 158-162.

15. Niewiadomska-Szynkiewicz, E., & Arabas, P. (2018). Resource management system for HPC computing. *In Conference on Automation* (pp. 52-61). Springer, Cham.
16. Nikolaou, P., Sazeides, Y., Lampropoulos, A., Guilhot, D., Bartoli, A., Papadimitriou, G., ... & Karakonstantis, G. (2019). On the Evaluation of the Total-Cost-of-Ownership Trade-offs in Edge vs Cloud deployments: A Wireless-Denial-of-Service Case Study. *IEEE Transactions on Sustainable Computing*.
17. Saad, Y. (2003). *Iterative methods for sparse linear systems*. 2nd ed. Society for Industrial and Applied Mathematics. 528
18. Singh, A. K., Dziurzanski, P., & Indrusiak, L. S. (2015). Value and energy optimizing dynamic resource allocation in many-core HPC systems. *In 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)* (pp. 180-185). IEEE.
19. Tay, S. I., Lee, T. C., Hamid, N. Z. A., & Ahmad, A. N. A. (2018). An overview of industry 4.0: Definition, components, and government initiatives. *Journal of Advanced Research in Dynamical and Control Systems, 10*(14), 1379-1387.
20. Zhang, R., & Zhu, Q. (2019). Consensus-based distributed discrete optimal transport for decentralized resource matching. *IEEE Transactions on Signal and Information Processing over Networks, 5*(3), 511–524.

References in Cyrillics

1. Borovkov, A.I. (2018). Smart technology serving product programs. *Project bulletin*, (2), 32-36.
2. Golshein, E.G., Borisova, E.P., & Dubson, M.S. (1990). Dialogue System for Analysis of Multi-Criteria Problems. Economics and mathematical methods. 26, 4.
3. Kozyrev, A.N. (1975). Optimization for resources distribution in a system of linear production models. *Optimization, 16*(33), 62.

Сергей Юрьевич Балычев (0000-0003-0162-232),
Финансовый университет при Правительстве РФ, Москва, Россия.
e-mail: bs0209@inbox.ru

Александр Михайлович Батьковский (0000-0002-5145-5748),
Центральный экономико-математический институт РАН, Москва, Россия
e-mail: batkovsky@yandex.ru

Михаил Александрович Батьковский (0000-0002-4930-0675),
АО «НИЦ «Интелэлектрон», Москва, Россия
e-mail: batkovsky@yandex.ru

Фёдор Анатольевич Белоусов (0000-0002-3040-3148),
Центральный экономико-математический институт РАН, Москва, Россия
e-mail: sky_tt@list.ru

Иван Викторович Неволин (0000-0002-8462-9011)
Центральный экономико-математический институт РАН, Москва, Россия
e-mail: i.nevolin@cemi.rssi.ru

Ключевые слова
Оптимизация, Высокопроизводительные вычисления, большие данные, нагрузка.

Sergey Yu Balychev, Aleksandr M. Batkovskiy, Michael A. Batkovskiy, Fedor A. Belousov, Ivan V. Nevolin, Shared Resources in Industry 4.0

Keywords
Optimization, High Performance Computing, Big data, load.

DOI: 10.34706/DE-2023-03-11
JEL classification: C61 – методы оптимизации, модели программирования, динамический анализ

Abstract
The purpose of the present study is to analyze the distribution of computing resources used by firms performing their activities based on the concept of Industry 4.0. A different level of production and management organization, increasing flexibility and adaptability of business systems through the development of automation and data exchange in constant interaction with the external environment – these aspects of the digital economy affect the complexity of the problem in context. The development of information technology should be accompanied by the development of new solutions to support the changes the industry faces. Therefore,

optimization of computing resources distribution given the spreading Industry 4.0 technologies becomes of high scientific and practical importance. The methodological approach used in the study grounds on fundamental results from research into information and communication technologies transformation, general principles of management theory, methods of economic analysis, mathematical modeling in economics. As a result of the study, a procedure to distribute computing resources within a firm is developed. The procedure is relevant in the context of Industry 4.0 and reflects the author's approach to the analysis of the process under consideration. The novelty of the results obtained consists of the developed algorithm that supports computing resources assignment to individual tasks under incomplete information. This algorithm sets a rather general framework for resources management and takes into account security issues. Given this, the algorithm described deserves attention in designing computing infrastructure of a firm. Implementation of the algorithm promises an increase in the efficiency of firms employing Industry 4.0 technologies and, accordingly, innovative development acceleration.